Combining Lexical-Syntactic Information with Machine Learning for Recognizing Textual Entailment

University of Jaén
Spain

A. Montejo-Ráez, J.M. Perea, F. Martínez-Santiago, M.A. García-Cumbreras, M. Martín-Valdivia, A.L. Ureña-López
SINAI Research Group – Dpto. de Informática, University of Jaén, 23071 – Jaén (Spain)

sistemas inteligentes de acceso a la información

Machine-Learning Approach

What is *Textual-Entailment?*

Given a text *t* and an hypothesis *h* we want to define a function *e* which takes these two texts as arguments and returns an answer to the entailment question *is* h *implicit in* t*?*

e(t,h) = { *YES if* h *is implicit in* t ; *NO otherwise* }

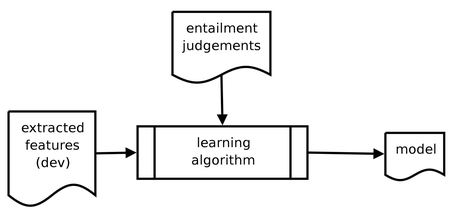
This can be approximated using Machine Learning algorithms by

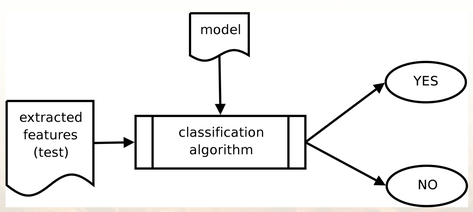
$$\hat{e}(t,h) = bc(f,m)$$

where

bc is a binary classifier
f is a set of features
m is the learned model of the classifier *bc*

Therefore, *training* has to be performed before solving entailment questions which are, with this approach, reduced to classification decisions:





Computed features

Lexical similarity

***SIM*: Simple Matching**

Semantic distance between each stem of the hypothesis and the text. When the distance is under a certain threshold, then both stems are counted as similar.

$$SIM_{matching} = \frac{\sum_{i \in H} similarity(i)}{|H|}$$

$$similarity(i) = \begin{cases} 1 & if\ \exists j \in T, s_L(i,j) > 0.5 \\ 0 & otherwise \end{cases}$$

where
s_L is the Lin's similarity measure
H and *T* are the hypothesis and text sets of concepts respect.
i,j are two concepts (WordNet's synsets)

***BM*: Binary Matching**

Same as previous but:

$$similarity(i) = \begin{cases} 1 & if\ \exists j \in T, i = j \\ 0 & otherwise \end{cases}$$

***CSS*: Consecutive Subsequence Matching**

Counts the number of consecutive subsequences of stems that appear in both pieces of text.

***Trigrams*:**

Same as before, but using words instead of stems and only on subsequences of three words.

Syntactic similarity (*SynTree* features)

We computed a set of measures from the syntactic trees obtained with COLLINS from both texts. *Aligned* terms are identified (those stems appearing in both texts), and then, also with the POS information of the parsing, we compute:

1. Number of aligned terms
2. Number of coincident POS of aligned terms
3. Number of unmatched POS of aligned terms
4. Minimal, maximal and average of distances differences in nodes through the syntactic trees to go from one aligned term to another

Experimental results

We have performed 6 different experiments, applying combinations of described features and two possible learning algorithms: BBR and TiMBL

SIM	BM	CSS	Trigrams	SynTree	Classifier	Accuracy
X		X	X	X	BBR	0.6475
X	X	X	X	X	BBR	0.6462
X		X	X		BBR	0.6387
X	X	X	X		TiMBL	0.6062
X		X	X		TiMBL	0.6037
X		X	X	X	TiMBL	0.5700

Conclusions and future work

Good results areobtained integrating all features, but lexical ones seem to hold better information, being the most crucial feature the simplest one: *trigams* weighting.

We plan to combine the output of several classifiers (like also SVM) to produce a single answer. Also the syntactic information has to be studied in deeper detail. Besides, a model working at semantic level is under implementation.

References

- Alexander Budanitsky and Graeme Hirst. 2001. *Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures*.
- Michael Collins. 1999. *Head-Driven Statistical Models for Natural Language Parsing*. Ph.D. Thesis, University of Pennsylvania
- Walter Daelemans, Jakub Zavrel, Ko van der Sloot and Antal va den Bosc.1998. *TiMBL: Tilburg memory based learner*, version 1.0
- Oscar Ferrandez, Daniel Micolo, Rafael Muñoz and Manuel Palomar. 2007. *Técnicas léxico-sintácticas para reconocimiento de implicación textual*. Tecnologías de la Información Multilingüe y Multimodal (*in press)*
- Jay J. Jiang and David W. Conrath. 1997. *Semantic similarity based on corpus statistics and lexical taxonomy*. In Proceedings of International Confeence on Research in Computational Linguistics. Taiwan.
- Dekang Lin. 1998. *An information-theoretic definition of similarity*. In Proceedings of the 15th International Conference on Machine Learning.
- Philip Resnik. 1995. *Using information content to evaluate semantic similarity*. In Proceedings of the 14thInternational Joint Conference on Artificial Intelligence, Montreal.